PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : H04L 12/56, 25/14 — A1

(11) International Publication Number: WO 93/19550

(43) International Publication Date: 30 September 1993 (30.09.93)

(21) International Application Number: PCT/SE93/00227

(22) International Filing Date: 16 March 1993 (16.03.93)

(30) Priority data:
9200813-5 17 March 1992 (17.03.92) SE

(71) Applicant: TELEFONAKTIEBOLAGET LM ERICSSON [SE/SE]; S-126 25 Stockholm (SE).

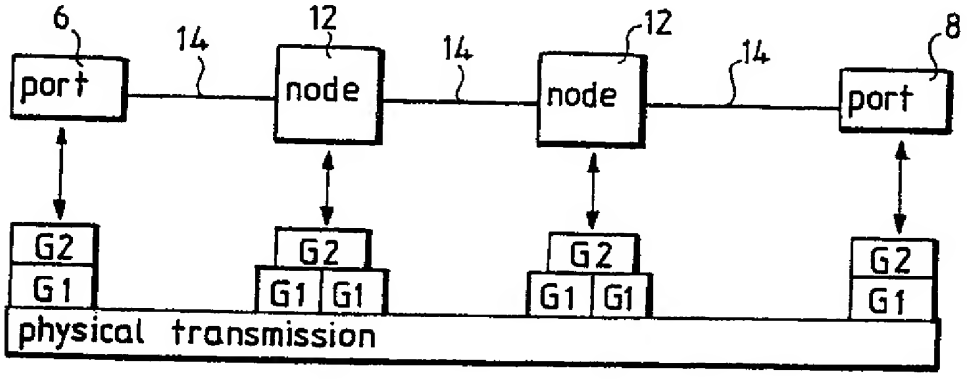
(72) Inventors: ERIKSSON, Thomas, Agne ; Varvsgatan 8 B, S-117 29 Stockholm (SE). JIANG, Hao ; Lillhagsvägen 30, S-124 71 Bandhagen (SE). LJUNGBERG, Per, Arvid, Martin ; Rödabergsgatan 4, S-113 33 Stockholm (SE). SANDIN, Rolf, Stefan ; Skogstorpsvägen 40, S-191 39 Sollentuna (SE).

(74) Agents: DELHAGE, Einar et al.; Bergenstråhle & Lindvall AB, Box 17704, S-118 93 Stockholm (SE).

(81) Designated States: AU, BR, CA, FI, JP, KR, NO, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: A METHOD OF GROUPING LINKS IN A PACKET SWITCH



(57) Abstract

In a packet switch, which is intended for packets of constant length, and which, between switch ports, is divided into a number of nodes (12) and transmission links (14), where the nodes carry out space selection and the transmission links offer point-to-point transmission between the nodes, link grouping is carried out. More particularly, from multiple parallel physical links coming into the switch, link groups are produced, each in the form of a logical link with a bandwidth that is the sum of the bandwidths of the physical links included in the link group, said logical link being restored to outgoing parallel physical links from the switch. One or more link protocols (G1, G2) are used according to which a label in the packet's header is made to describe a route that holds together the grouped links through the entire switch, such that bits in the label describing the route over a certain transmission link are the same for packets belonging to the same link group.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**FR**	France	**MR**	Mauritania
AU	Australia	**GA**	Gabon	**MW**	Malawi
BB	Barbados	**GB**	United Kingdom	**NL**	Netherlands
BE	Belgium	**GN**	Guinea	**NO**	Norway
BF	Burkina Faso	**GR**	Greece	**NZ**	New Zealand
BG	Bulgaria	**HU**	Hungary	**PL**	Poland
BJ	Benin	**IE**	Ireland	**PT**	Portugal
BR	Brazil	**IT**	Italy	**RO**	Romania
CA	Canada	**JP**	Japan	**RU**	Russian Federation
CF	Central African Republic	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CG	Congo			**SE**	Sweden
CH	Switzerland	**KR**	Republic of Korea	**SK**	Slovak Republic
CI	Côte d'Ivoire	**KZ**	Kazakhstan	**SN**	Senegal
CM	Cameroon	**LI**	Liechtenstein	**SU**	Soviet Union
CS	Czechoslovakia	**LK**	Sri Lanka	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**MC**	Monaco	**UA**	Ukraine
DK	Denmark	**MG**	Madagascar	**US**	United States of America
ES	Spain	**ML**	Mali	**VN**	Viet Nam
FI	Finland	**MN**	Mongolia		

A METHOD OF GROUPING LINKS IN A PACKET SWITCH.

TECHNICAL BACKGROUND.

The present invention relates in general to link grouping in packet switches for packets with constant length, e.g., ATM switches (ATM = Asynchronous Transfer Mode). Link grouping means that a logical link is created consisting of many parallel physical links. The logical link offers a bandwidth that is the sum of that of the physical links. In this way can, for example, a logical link with a bandwidth of 622.08 Mb/s be provided through link grouping of four physical 155.52 Mb/s links.

More particularly the invention relates to a method of grouping links in a packet switch for constant-length packets, the packet switch having a plurality of nodes and a plurality of transmission links for selectively connecting a plurality of input ports to a plurality of output ports via routes through the switch, wherein the nodes enable space selection and the transmission links interconnect the nodes.

In an ATM-switch for wideband telecommunication there is a need for higher bandwidth than that which is physically available. Link grouping can offer good traffic characteristics and hardware flexibility and effectivity, for example smaller delay, fewer buffers, and the realization of different types of concentrators/multiplexers and switches in different speed classes while using the same base components. Blocking characteristics in a link coupled network, for example a Clos-network, are also improved through logical speedup, achieved with the help of link grouping.

STATE OF THE ART.

Methods available thus far for offering a wideband link are:

1. high physical speed,
2. series-parallel conversion in order to attain higher

logical speed than the physical one.

Disadvantages with the first method are the difficulty in realizing high speed electronics and mechanics, as well as the consequent power dissipation. The other method builds normally on multiplexing at bit-level and therefore requires bit-synchronous transmission.

Among publications discussing the need to create greater bandwidth than that which is physically available can be mentioned EP 0 374 574, WO 85/04300, WO 90/12467, and IBM Technical Disclosure Bulletin Vol 32, No. 9A, February 1990, pp. 45-49.

STATEMENT OF THE INVENTION.

One object with the present invention is to provide a method for doing link grouping of the type presented by way of introduction, which offers a simpler solution than previously known for the problem in the ATM context of achieving higher bandwidth than that which is physically available.

The method according to the invention comprises the steps of:

grouping, according to a link protocol, a plurality of incoming physical links into a link group, wherein the incoming physical links are connected in parallel to respective input ports, the link group has a bandwidth that is a sum of bandwidths of the grouped incoming physical links, and the link protocol requires labels in packets communicated via the link group, the labels describing a route through the switch that holds together the grouped incoming physical links and each label comprising at least one bit that describes the route and is identical to corresponding bits in labels of other packets communicated via the link group; and

distributing packets communicated via the grouped incoming physical links among a plurality of outgoing physical links connected in parallel to output ports of the switch.

DESCRIPTION OF DRAWINGS.

The invention shall now be described in more detail with

reference to the attached drawings, on which

Figure 1 schematically illustrates the general principle of link grouping to create logical links in an ATM switch,

Figure 2 shows a protocol model illustrating how to provide, according to the invention, link grouping in switches,

Figure 3 schematically illustrates multi-port link grouping according to the invention,

Figure 4 similarly schematically illustrates multi-port link grouping in a shared buffer pool

Figure 5 similarly schematically illustrates transmission of an internal flow of packets in one block to another block via grouped links,

Figure 6 show a logical structure for link grouping over a link between nodes,

Figure 7 schematically illustrates label attachment in switch ports for interport link grouping according to the invention.

PREFERRED EMBODIMENTS.

The same reference numbers are used in the various Figures for the same or equivalent details.

In fig. 1, label 2 indicates a switch core in a packet switch for packets with constant length, e.g. an ATM switch. A number of parallel physical links 4 arrive at input switch ports 6 to the core 2. Before the core 2 occurs a link grouping of the physical links 4 for forming logical links, each consisting of several of the parallel links 4. This link grouping is indicated in fig. 1, as in the following Figures, with ovals, for example, the ovals 1 and m in fig. 1. The link grouping at the core's 2 outputs need not necessarily be the same as at the inputs, which is indicated in fig. 1 where link groupings n and r at the outputs correspond to link groupings 1 and m, respectively, at the inputs.

Via output switch ports 8 a number of parallel, physical links 10 then leave the switch.

The link grouping is characterized more particularly in that the individual links in a group carry serial packet flows

at the the interface of the switch core but constitute one logical link by use of protocols (supported by hardware) which are specific for a certain grouping definition. The link grouping should be as general as possible, that is, grouping in many different group sizes should be possible, as shown in fig. 1. However, it is in practice most interesting to have the same link grouping at inputs and outputs, and to have groups which are multiples of four physical links.

The link grouping can be described using a protocol stack of the type shown in fig. 2. The upper part of this Figure indicates schematically how the switch core 2 can be divided into a number of nodes 12 and transmission links 14, where the nodes 12 perform space selection and the transmission links 14 offer point-to-point transmission between the nodes.

Between the switch ports 6 and the nodes 12 the link grouping is defined by a protocol on level 2, G2, and over the transmission links a protocol on level 1, G1, is used. The association between elements in the upper and lower parts of fig. 2 is indicated by double arrows.

That which is common for all thinkable variations of link grouping according to the invention is that the label number in the head of the packets describes a route that holds together the grouped links through the entire switch, which means that the bits in the label which describe the route over a certain transmission link are the same for packets which belong to the same group. This information is a part of G2.

Depending on how the rest of the protocol looks in level 2 (G2), link grouping can be divided into two types, namely multiport and interport link grouping.

Multiport link grouping shall be discussed first here with reference to Figures 3 to 6.

The advantage to this type of link grouping is that no extra functions are required in the switch ports, and its mechanism can more particularly be divided into what happens partly in the nodes and partly over the transmission links.

In a node the group is held together by using similar space selection for the grouped links, for example with the help of

certain bits in the label number being identical, so that they come out on the same grouped output and the sequence is retained. This can be realized, for example, according to what is shown in fig. 3, by links belonging to the same group, or logical links, for example n or m, being written into the same buffer 16 and 18, respectively, in the node 12, and being read out from the node serially on out-going links.

The common buffer principle, indicated with 20 in fig. 4, can otherwise also be used, wherein link grouping and readout can be controlled by a function 22 for memory handling according to known principles, see for example IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, October 1991, volume 9, number 8, p. 1239, "32*32 Shared Buffer Type ATM Switch VLSI's for B-ISDN's". In fig. 4 out-going logical links have been labeled with 24 and 26.

The functionality which must be introduced over a link is some type of simple link protocol (G1) in order to send and re-create a packet flow over a "parallel connection" with a division on the packet level according to the description below.

With reference to fig. 5, a flow 28 of packets 30 with constant length internally in a block can be transmitted to another block via external links 32. The internal flow can be said to have a packet frequency of F packets/s. It follows that the time for a packet in the internal flow can be said to be t=1/F.

The external links 32 can because of technical limits have a transmission capacity which is limited such that an external link can not handle a packet frequency of F, but only lower packet frequencies. In such a case can several external links 32 together, through grouping, transmit the internal packet flow 28.

When n external links are grouped and the packets 30 in the internal flow 28 are divided among them the packet frequency on every external link becomes F/n. It follows that the time for every packet on the external links becomes nt.

The sequence order between the packets is important and

must be preserved. It is therefore important that the packets 30 are transmitted on the grouped links 32 in such a way that the sequence ordering can be preserved. In order for this to be possible grouping logic 34 is needed on the sending side, and restoration logic 36 on the receiving side.

The send and receive logic, 34 and 36 respectively, should be designed so that a certain spread in transmission time is allowed between the external links 32.

The link protocol includes firstly an algorithm for how the packets shall be portioned out and gathered in between the physical links, such that space correctness is preserved, for example in the sequence 1, 2, 3, 4.

Secondly, the protocol includes synchronization which secures that the packets are transmitted such that there exists a determined time pattern between the physical links. For example the packets can be sent out simultaneously or time delayed by a constant interval. The time delay/difference in length on the links, can be allowed to correspond to one packet if suitable bit and packet synchronization is introduced.

First a short description shall be given here of a suitable design for the sending and receiving logic.

The design should be done so that the packets 30 in the internal flow 28 are divided over the external links 32 according to a given pattern. One suitable pattern is a cyclical one according to the following: Packet 1 is sent on external link 32.1, packet 2 on link 32.2, etc., until packet n on link 32.n; thereafter packet n+1 on link 32.1, etc., according to fig. 5. The packets on the external links should be sent with a mutual time delay such that if a packet (the begin of packets) is sent at time 0, packet 2 is sent at time t, etc., until packet n is sent at time (n-1)t, and thereafter packet n+1 at time nt.

Since the division is done at packet level, two buffer stages, 38 and 39, must be introduced according to fig. 6 because the bandwidth in to and out of the send queue is higher than the bandwidth of the physical links. The buffer stages compensate for the difference in bandwidth.

The grouping logic is built consisting of a distributing unit 42, one of the buffers 38.1 to 38.n per external link in the buffer stage 38, and a sequence control unit 44 that controls the time ordering between the packets on the external links. The distributing unit 42 allocates the packets, packet for packet, to the buffers 38.1 to 38.n toward the external links such that the distribution of packets between the links is obtained as described above.

The buffers 38.1 to 38.n, one to every external link, handles the differences in time between transmission of a packet on the internal flow verses on the external links. A buffer is filled with a packet from the internal flow in the time t. In the next time period of nt it is emptied toward the external link. To secure that the packet sequence is maintained and no packets pass by each other in the buffers, every buffer should have room for more than 1-1/n packets and less than 2-1/n packets.

The unit 44 for controlling the time ordering sees to it that the above indicated time ordering is maintained. The unit uses the packet flow on external link 32.1 as reference and controls the flows out of the buffers 32.2 to 32.n toward the other links based on this reference. Since the packet frequency is the same for all the external links the unit only needs to be activated upon initial startup or when some external link is restarted.

The restoration logic 36 consists of one of the buffers 40.1 to 40.n per external link, a gathering unit 46, and a sequence restoration unit 48.

Differences in transmission times between the different external links can arise. These differences can arise in the grouping logic 34, in the transport over the external links, as well as in the restoration logic 36. The restoration logic must therefore be built such that it can handle a spread in arrival time for the packets on the external links.

External link 32.1 is used as the reference. On the other external links the nominal arrival times for the packets are calculated starting from the reference. From the nominal

arrival time a spread of +d is allowed. Normally d can be allowed to be d=nt/2, i.e., a half packet time on the grouped links.

The buffers 40 to the external links 32 are used to handle the differences in packet time between the external links and the internal flow. A buffer is filled with a packet from the external link during the time nt. The buffer is emptied thereafter toward the internal flow in the time t. In the buffers 40 the packet is also stored for an extra time of nominally d and maximum 2d to handle differences in actual verses nominal arrival time for packets on different links. The buffer needs to be maximum 2 packets long.

The gathering unit 46 picks up packets from the buffers 40 to the external links according to the order that is described above.

The sequence restoring unit 48 sees to it that packets are picked up in the order that is described above.

With interport link grouping the protocols act only between the switch ports. Two methods are conceivable here.

The first method is label based. With reference to fig. 7, a label, based on VCI/VPI (Virtual Circuit Identifier/ Virtual Path Identifier) information, is attached in the switch ports 6 to arriving packets in sequence for the grouped links. The switch core does not need to "know" anything about which physical links belong together. Only level 2 protocol is needed. The packets are supplied suitably with sequence numbers to be able to restore the sequence order if packets within the same group are delayed by different amounts of time along different physical routes.

With the above mentioned label attachment in the switch ports 6 a buffer is used that has a number of outputs corresponding to the number of links to be grouped. The links are written into the buffer and are read out sequentially on the outputs. The read-out routine is preferably arranged so that the incoming packets are distributed evenly over all the links in the group. The grouped packets also receive a label code that essentially contains the same bits. The grouping

means must be able to buffer enough packets to be able, with help of the labels, to choose packets that shall go out on the same logical link, correct the sequence, and evenly distribute them on all the links in the group. Odd numbers of packets are filled out with empty packets. The link terminator 8 also contains a buffer large enough to enable correction of the sequence.

The other method is based on packet synchronism through the switch core. Protocols on both level 1 and level 2 are used.

The protocol on level 2 is based on a label attacher and a link terminator as in method one. The difference is that no sequence number is needed to secure the sequence, nor is any large buffer needed. On level 1 packet synchronism is secured in and between the switch ports and the nodes, so that the packets in the group move through the entire switch packet-synchronously. Sorting in the switch ports is carried through with help of the bits in the label that are different between packets within the group.

CLAIMS.

1. A method of grouping links in a packet switch for constant-length packets, the packet switch having a plurality of nodes and a plurality of transmission links for selectively connecting a plurality of input ports to a plurality of output ports via routes through the switch, wherein the nodes enable space selection and the transmission links interconnect the nodes, comprising the steps of:

grouping, according to a link protocol, a plurality of incoming physical links into a link group, wherein the incoming physical links are connected in parallel to respective input ports, the link group has a bandwidth that is a sum of bandwidths of the grouped incoming physical links, and the link protocol requires labels in packets communicated via the link group, the labels describing a route through the switch that holds together the grouped incoming physical links and each label comprising at least one bit that describes the route and is identical to corresponding bits in labels of other packets communicated via the link group; and

distributing packets communicated via the grouped incoming physical links among a plurality of outgoing physical links connected in parallel to output ports of the switch.

2. The method of claim 1, wherein the link group is held together by space selection in the nodes based on the labels such that a sequence of the grouped incoming physical links are connected to a corresponding sequence of outgoing physical links.

3. The method of claim 2, wherein each node includes a plurality of buffer memories, the grouping step includes the step of storing packets communicated via the link group in the same buffer memory in each node, and the distributing step includes the steps of retrieving the stored packets from the buffer memories and providing the retrieved packets to the outgoing physical links.

4. The method of claim 2, wherein each node includes a buffer memory, the grouping step includes the step of storing packets communicated via a plurality of link groups in the

buffer memory in each node, and the distributing step includes the steps of retrieving the stored packets from the buffer memories and providing the retrieved packets to respective groups of outgoing physical links.

5. The method of any of claims 2 to 4, wherein the link protocol provides for properly sequencing packets communicated via the link group and synchronizing such packets based on a predetermined timing pattern.

6. The method of claim 5, wherein the link protocol provides for one of simultaneously transmitting the packets communicated via the link group and transmitting the packets communicated via the link group with a predetermined time delay between such packets.

7. The method of claim 1, wherein the link protocol acts on packets only between the input ports and output ports, providing for interport link grouping.

8. The method of claim 7, further comprising the steps of sequentially adding labels to packets arriving at the switch via the link group and sorting such packets based on the labels so that an arrival sequence of such packets is preserved.

9. The method of claim 7, wherein the switch includes a buffer memory having a plurality of outputs corresponding to the plurality of incoming physical links in the link group, the grouping step includes the step of storing packets communicated via the link group in the buffer memory, and the distributing step includes the steps of sequentially retrieving the stored packets from the buffer memory and providing the retrieved packets to the buffer's outputs.

10. The method of claim 9, wherein the retrieved packets are provided evenly to the buffer's outputs.

11. The method of claim 10, wherein the labels in packets communicated via the link group each comprise a predetermined set of bits.

12. The method of claim 11, wherein a plurality of packets are stored in the buffer memory, the distributing step includes the steps of selecting packets to be retrieved for an outgoing link group comprising a plurality of outgoing physical links,

and the retrieved packets are provided evenly to the plurality of outgoing physical links.

13. The method of claim 12, wherein the distributing step includes the step of padding a plurality of packets comprising an odd number of packets with a predetermined number of empty packets.

14. The method of any of claims 9 to 13, including the step of synchronizing the packets during the grouping and distributing steps such that the packets flow synchronously through the switch.

15. The method of claim 14, wherein the grouping step includes the step of sorting the packets based on differences between bits in the packets' labels.

Fig.1

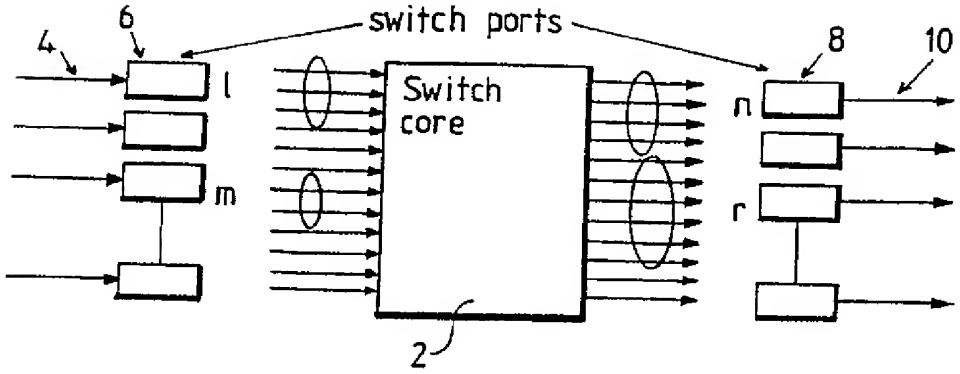


Fig.2

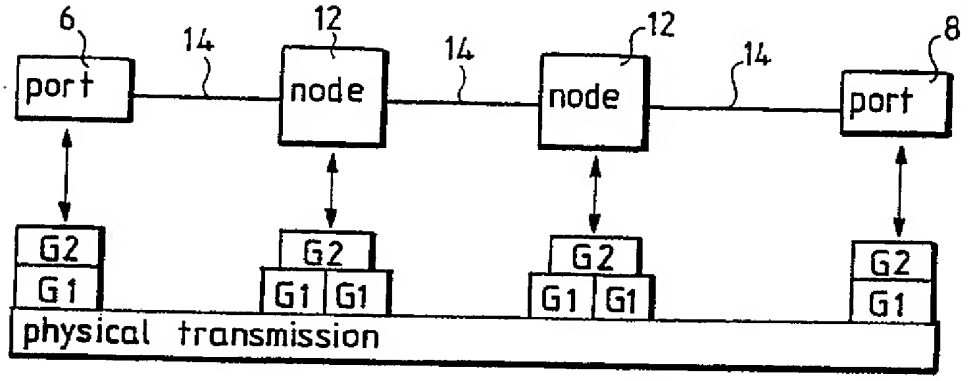


Fig.3

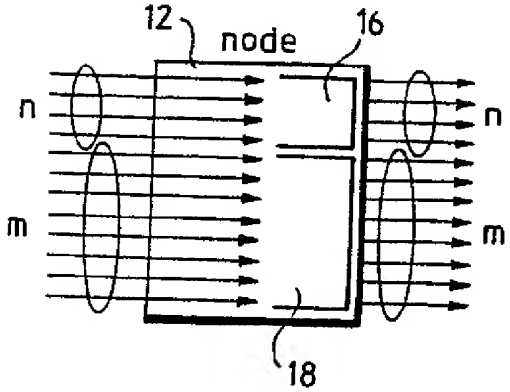


Fig.4

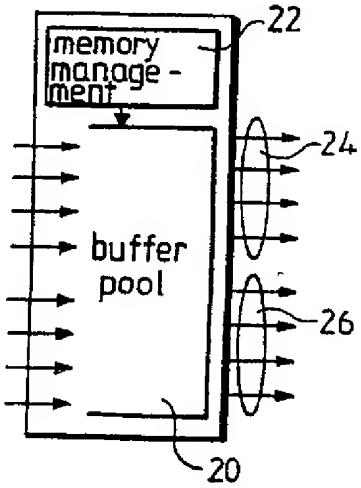


Fig. 5

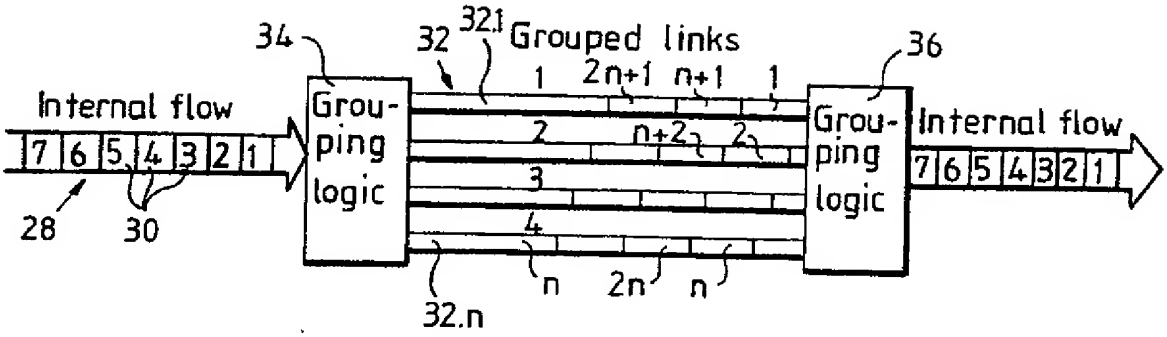


Fig. 6

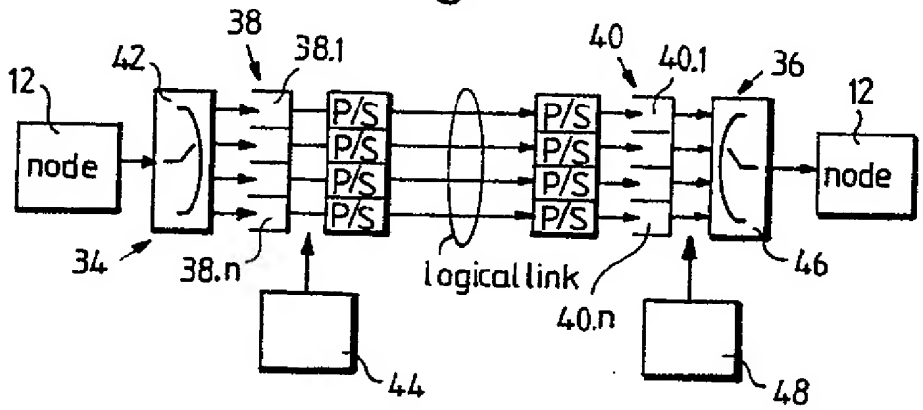
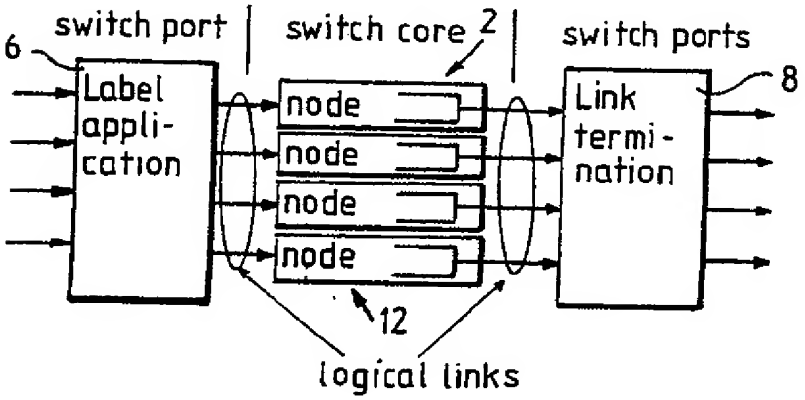


Fig. 7



INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 93/00227

A. CLASSIFICATION OF SUBJECT MATTER

IPC5: H04L 12/56, H04L 25/14
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC5: H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, Volume 9, October 1991, Takahiko Kozaki et al, "32 32 Shared Buffer Type ATM Switch VLSI's for B-ISDN's", page 1239 - page 1247, cited in the application --	1-15
A	IBM Technical Disclosure Bulletin, Volume 32, February 1990, . , "PARALLELING LINKS IN A UNIFIED SWITCHING ARCHITECTURE FOR CIRCUIT/PACKET SWITCHING", page 45 - page 49, cited in the application --	1-15
A	US, A, 4870641 (ACHILLE PATTAVINA), 26 Sept 1989 (26.09.89), abstract --	1-15

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 May 1993	21 -06- 1993
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Rune Bengtsson Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 93/00227

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP, A1, 0159810 (BRITISH TELECOMMUNICATIONS PLC.), 30 October 1985 (30.10.85), cited in the applications --	1-15
A	EP, A2, 0374578 (SIEMENS AKTIENGESELLSCHAFT), 27 June 1990 (27.06.90), abstract --	1-15
A	EP, A2, 0374574 (SIEMENS AKTIENGESELLSCHAFT), 27 June 1990 (27.06.90), cited in the applications --	1-15
A	WO, A1, 9012467 (HOFMAN-BANG & BOUTARD A/S), 18 October 1990 (18.10.90), cited in the application --	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

28/05/93

International application No.
PCT/SE 93/00227

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US-A- 4870641	26/09/89	EP-A- 0408648	23/01/91
EP-A1- 0159810	30/10/85	SE-T3- 0159810	
		AU-B- 575635	04/08/88
		AU-A- 4113585	11/10/85
		CA-A- 1277743	11/12/90
		JP-B- 4067824	29/10/92
		JP-T- 61501543	24/07/86
		US-A- 4775987	04/10/88
		WO-A- 8504300	26/09/85
EP-A2- 0374578	27/06/90	AU-B- 617231	21/11/91
		AU-A- 4723189	28/06/90
		CA-A- 2006393	23/06/90
		JP-A- 2224444	06/09/90
		US-A- 5016245	14/05/91
EP-A2- 0374574	27/06/90	AU-A- 4722989	28/06/90
		CA-A- 2006392	23/06/90
		JP-A- 2224550	06/09/90
WO-A1- 9012467	18/10/90	AU-A- 5440290	05/11/90
		EP-A- 0466785	22/01/92